@rmoff

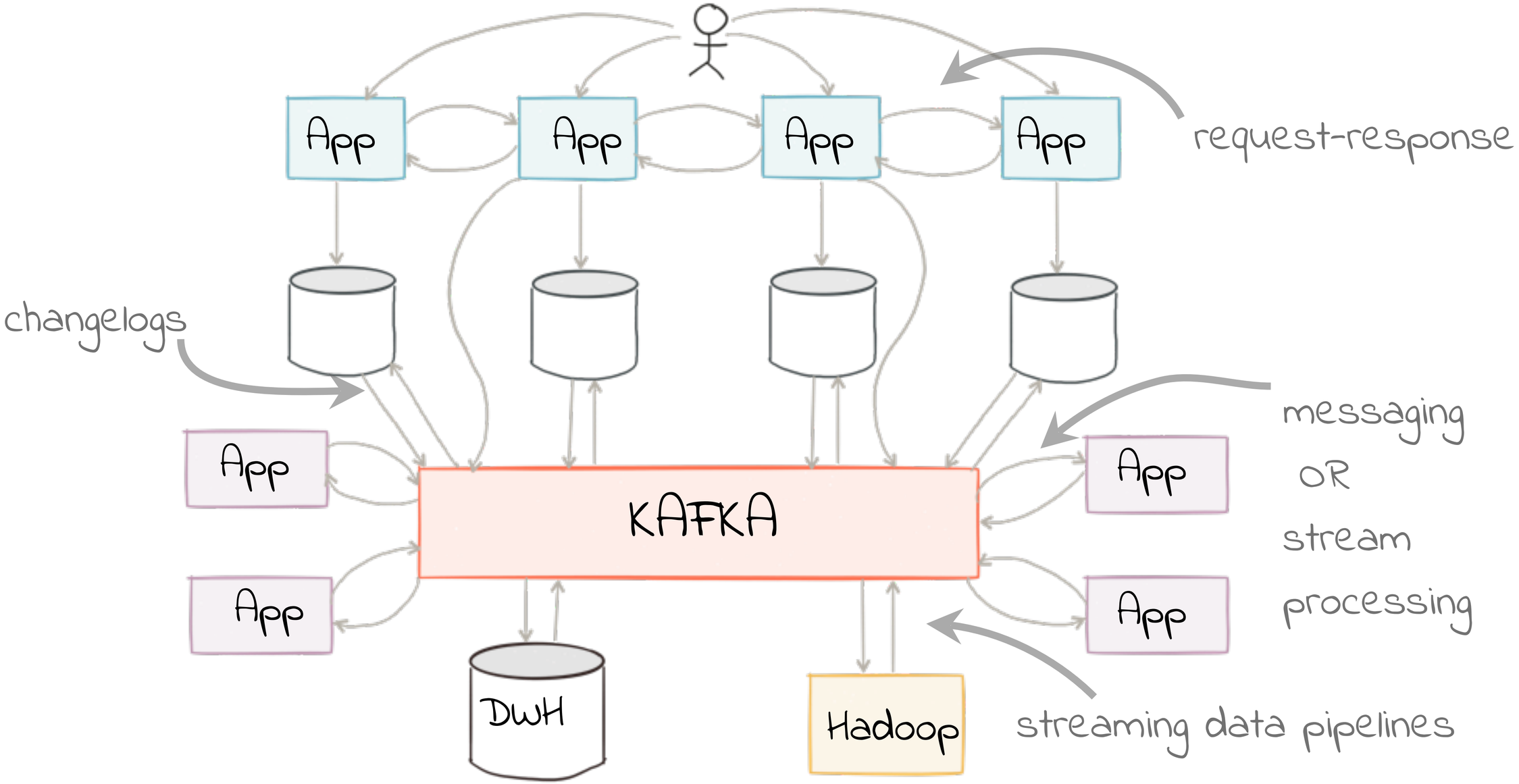
#ukoug_tech18

Integrating Databases and Apache Kafka

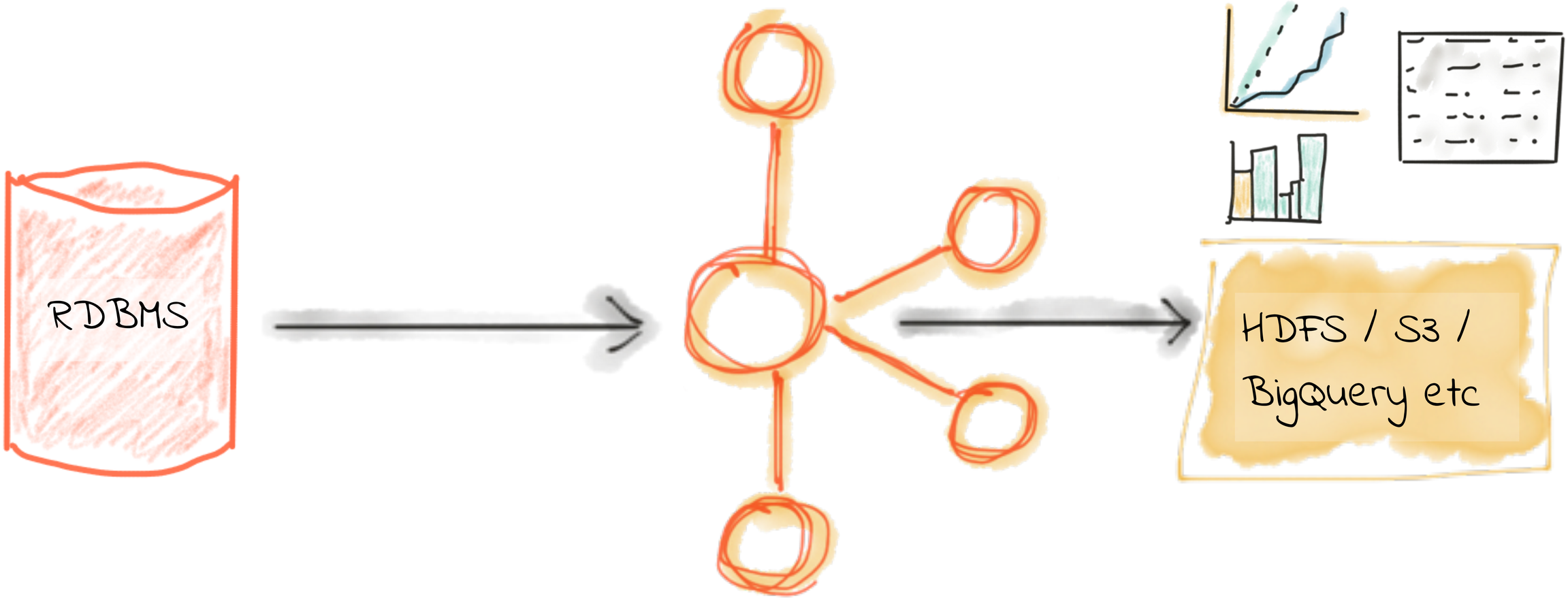
Robin Moffatt, Developer Advocate @ Confluent

Photo by Emily Morter on Unsplash

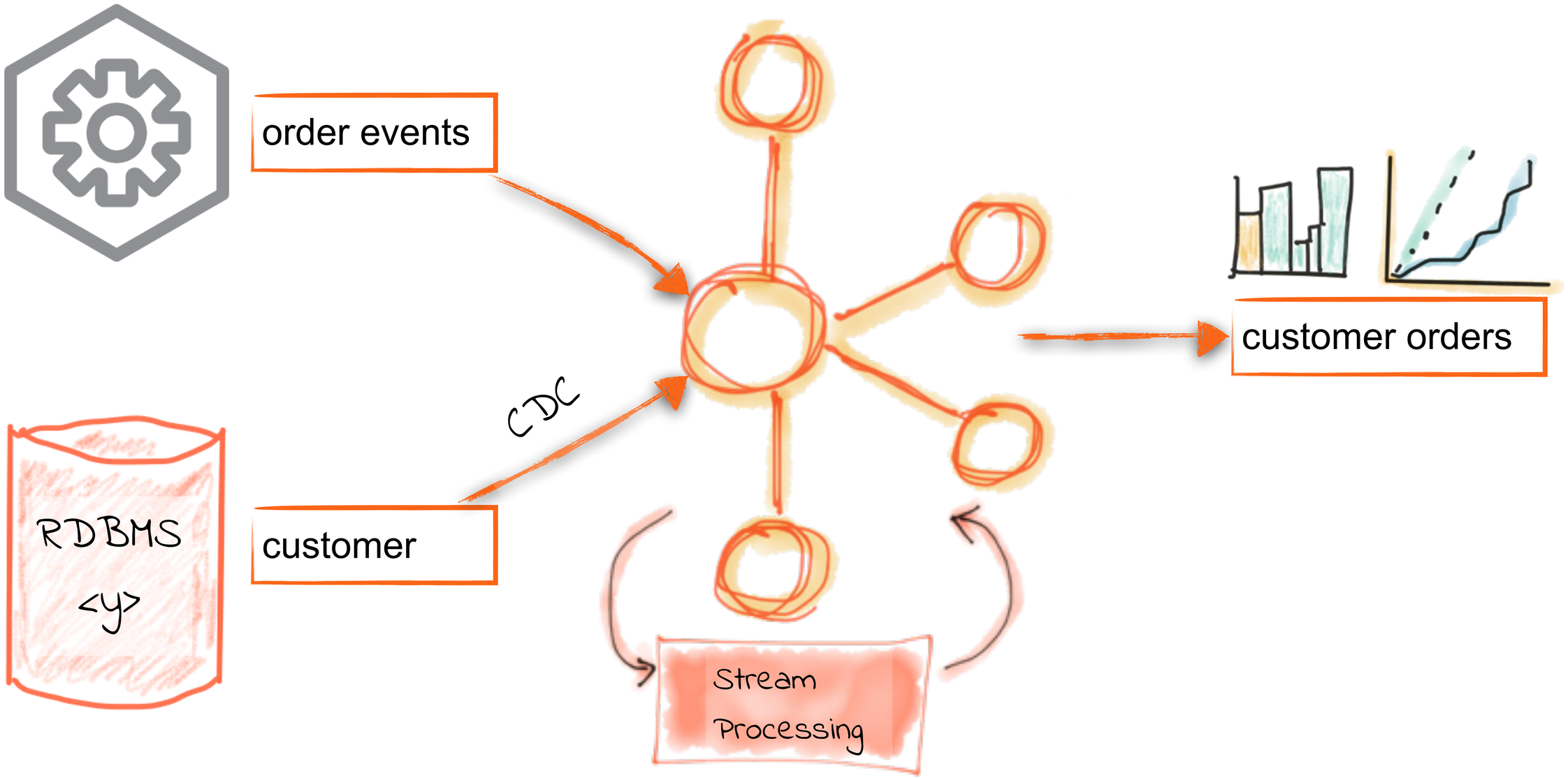
Kafka is a Streaming Platform



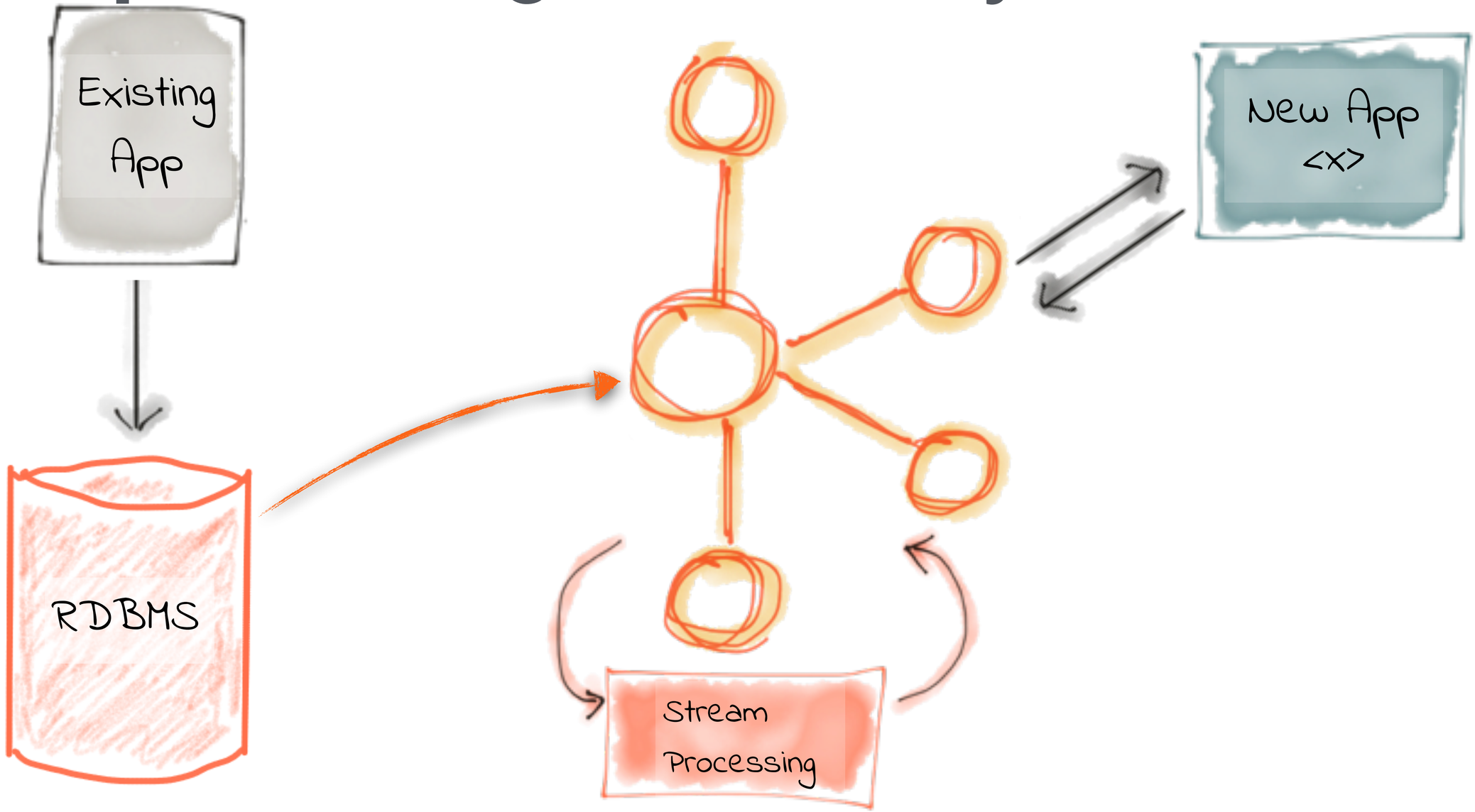
Analytics - Database Offload



Real-time Event Stream Enrichment



Evolve processing from old systems to new



"

But streaming...I've just got data in a database...right?

“

Bold claim: all your data is event streams

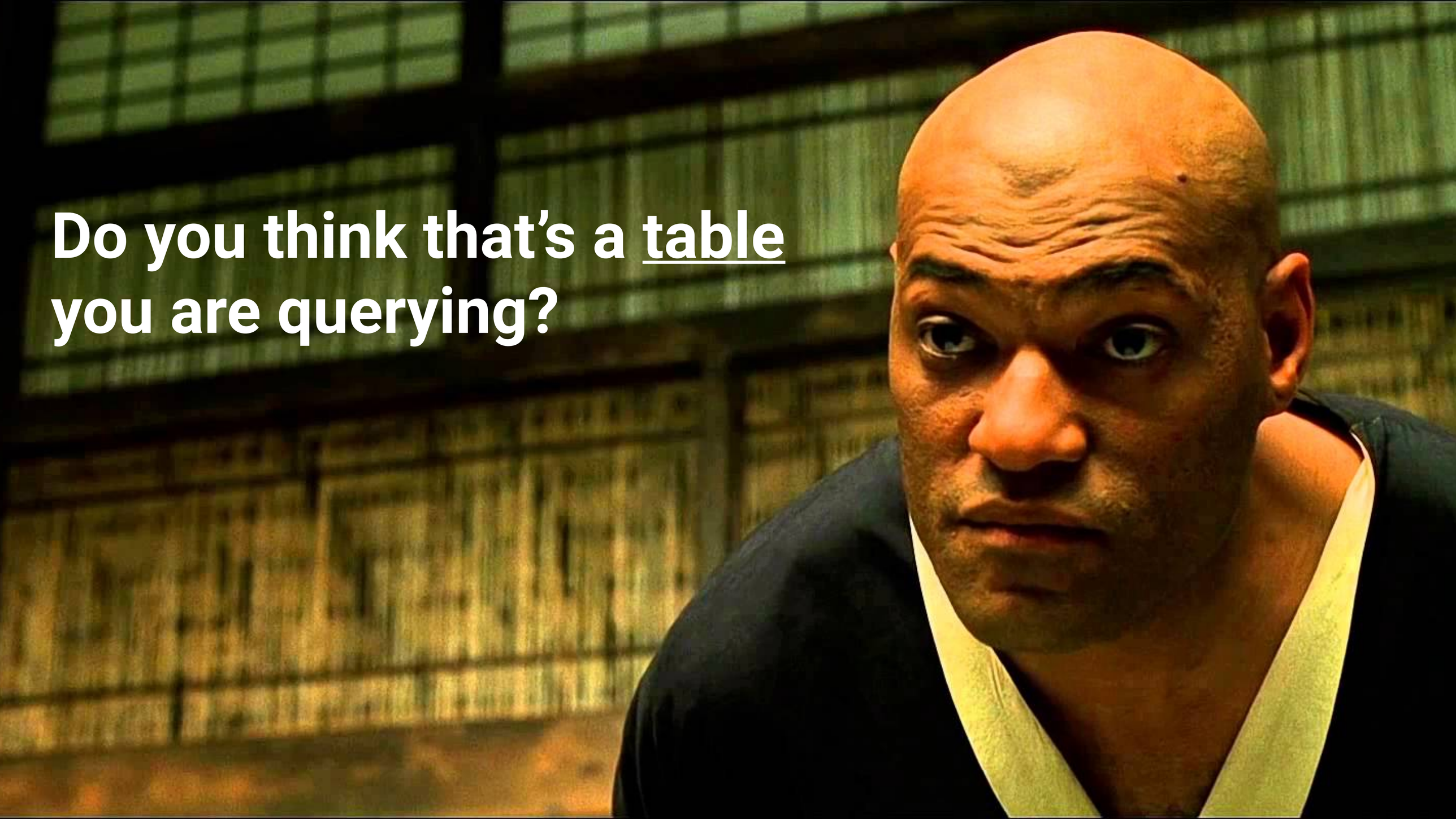
A Customer Experience

A Sale

A Sensor Reading

An Application Log Entry

Databases


Do you think that’s a table
you are querying?

The Stream Table Duality

Account ID	Balance
12345	€50

The Stream Table Duality

Account ID	Amount
12345	+ €50

Time

Account ID	Balance
12345	€50

The Stream Table Duality

Account ID	Amount
12345	+ €50
12345	+ €25

Time

Account ID	Balance
12345	€75

The Stream Table Duality

Time

Account ID	Amount
12345	+ €50
12345	+ €25
12345	-€60

Account ID	Balance
12345	€15

The Stream Table Duality

Stream

Time

Account ID	Amount
12345	+ €50
12345	+ €25
12345	-€60

Table

Account ID	Balance
12345	€15

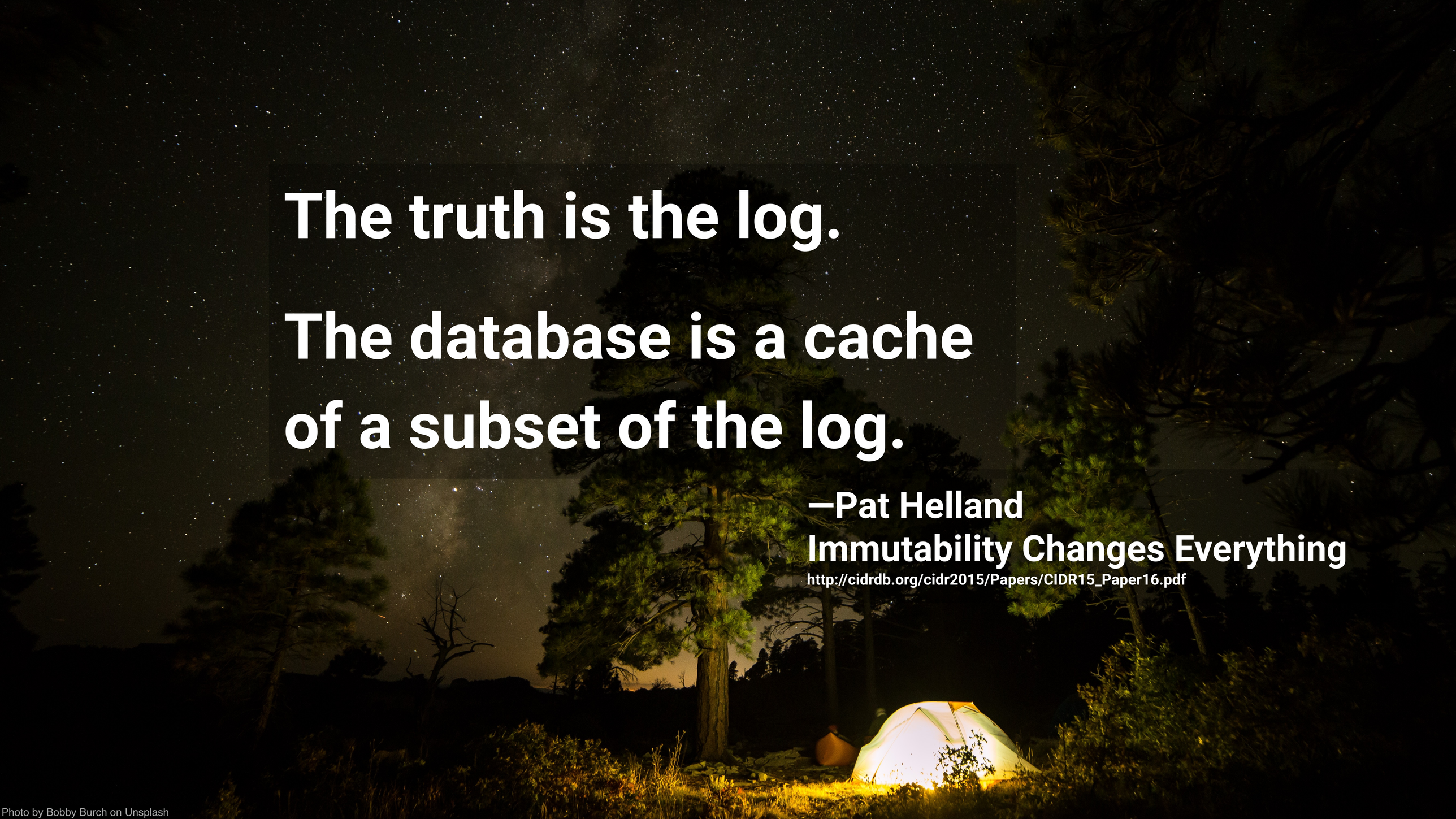
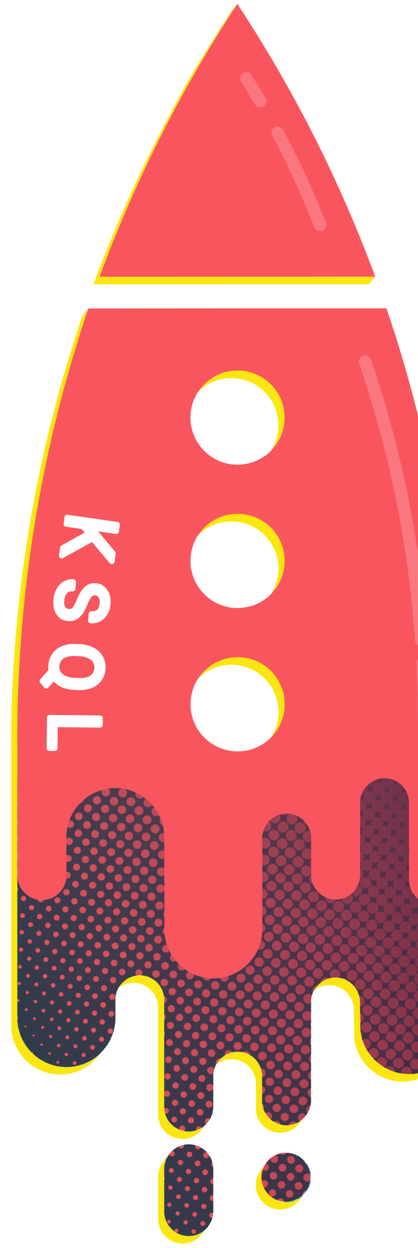


Photo by Bobby Burch on Unsplash

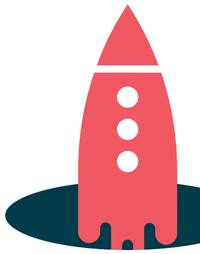
KSQL
is the
Streaming
SQL Engine
for
Apache Kafka



KSQL for Real-Time Monitoring

- Log data monitoring, tracking and alerting
- syslog data
- Sensor / IoT data

```
CREATE STREAM SYSLOG_INVALID_USERS AS
SELECT HOST, MESSAGE
FROM SYSLOG
WHERE MESSAGE LIKE '%Invalid user%';
```

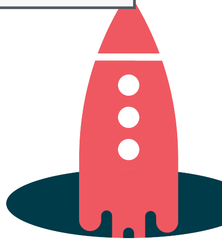


http://cnfl.io/syslogs-filtering / http://cnfl.io/syslog-alerting

KSQL for Streaming ETL

Joining, filtering, and aggregating streams of event data

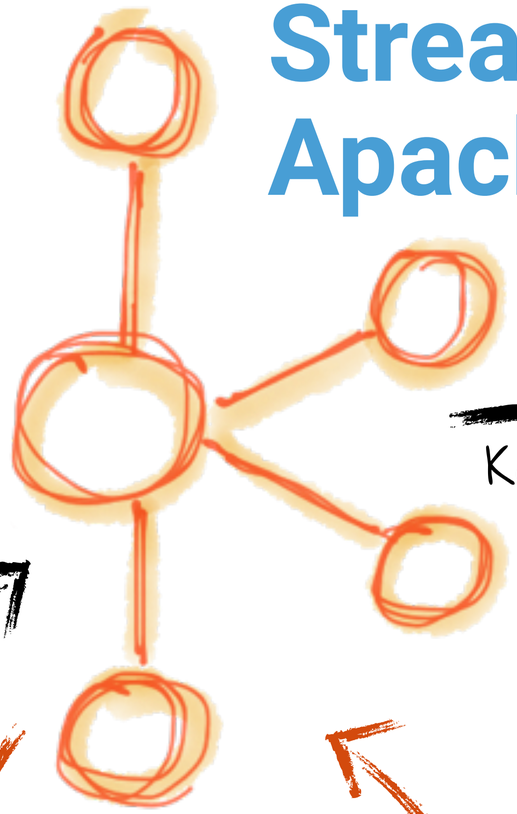
```
CREATE STREAM vip_actions AS
  SELECT userid, page, action
  FROM clickstream c
  LEFT JOIN users u
    ON c.userid = u.user_id
  WHERE u.level = 'Platinum';
```



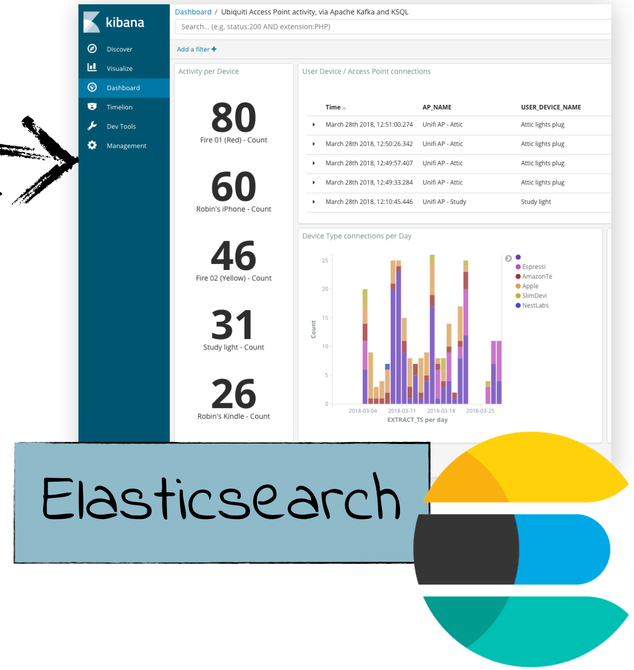
Streaming ETL with Apache Kafka and KSQL


```
{
  "rating_id": 2133,
  "user_id": 9,
  "stars": 1,
  "route_id": 7219,
  "rating_time": 1519402815063,
  "channel": "web",
  "message": "worst. flight. ever. #neveragain"
}
```


Producer API

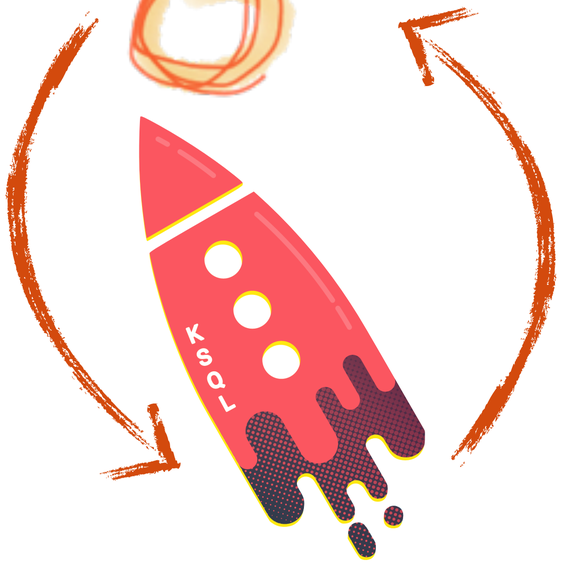


Kafka Connect



Elasticsearch

Kafka Connect
Debezium

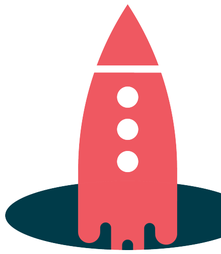



```
{
  "uid": 9,
  "name": "Neha",
  "locale": "en_US",
  "address_city": "Palo Alto",
  "elite": "P"
}
```

KSQL for Anomaly Detection

Identifying patterns or anomalies in real-time data, surfaced in milliseconds

```
CREATE TABLE possible_fraud AS
  SELECT card_number, count(*)
    FROM authorization_attempts
    WINDOW TUMBLING (SIZE 5 SECONDS)
    GROUP BY card_number
    HAVING count(*) > 3;
```



@rmoff #ukoug_tech18

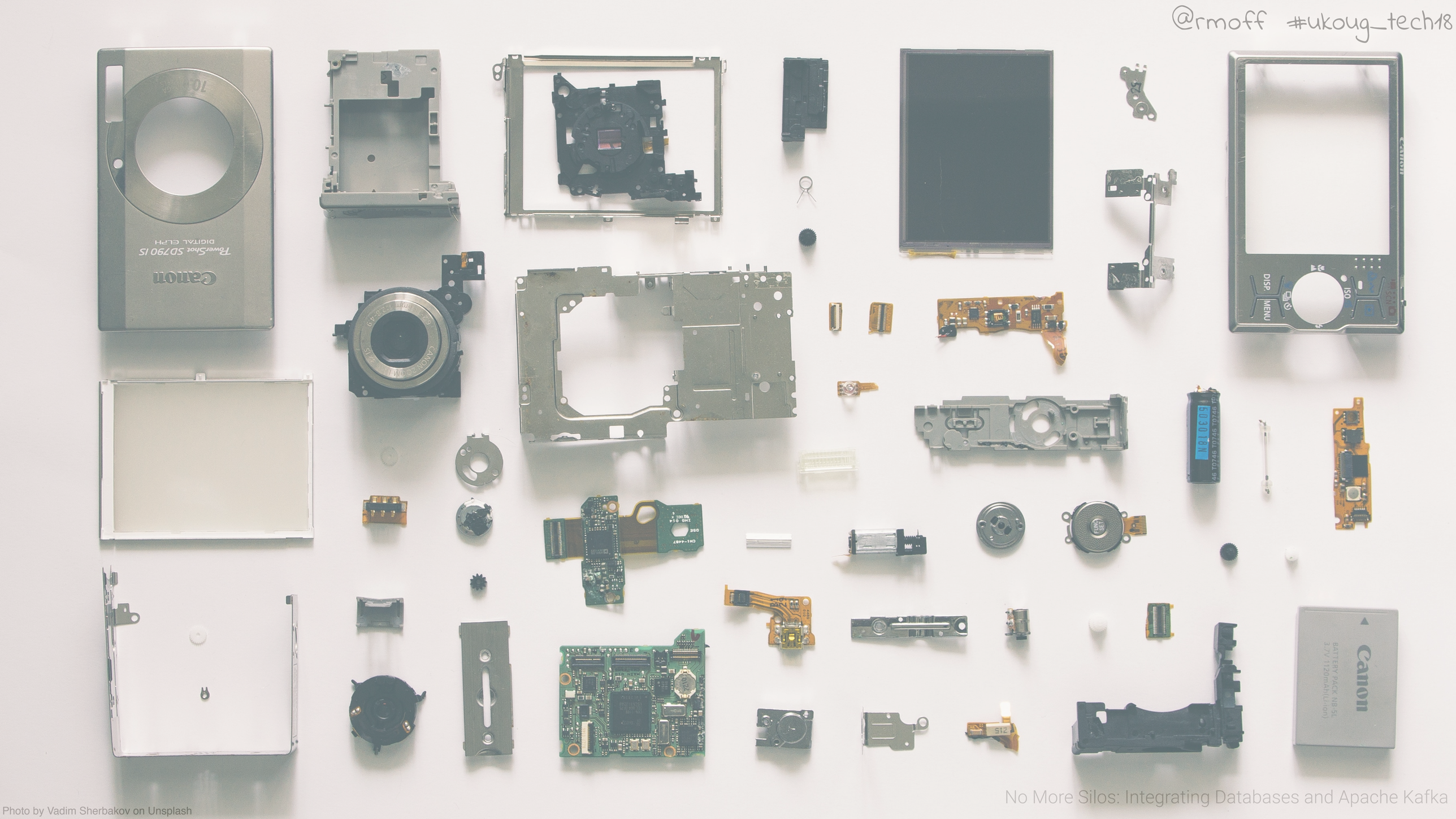
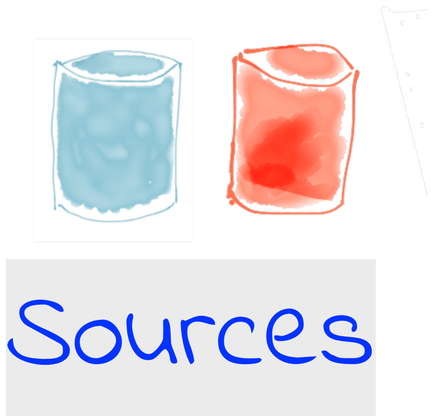


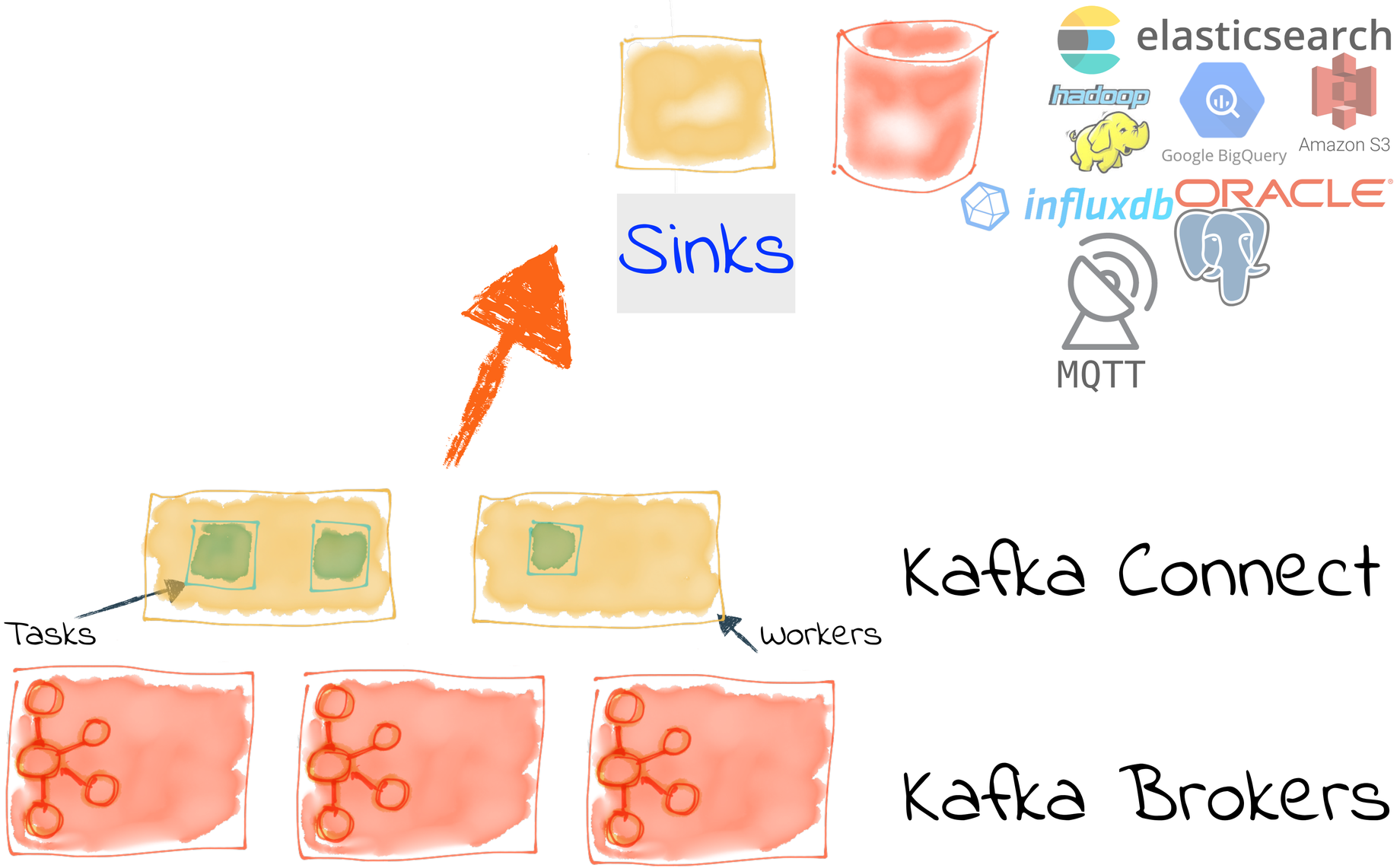
Photo by Vadim Sherbakov on Unsplash

No More Silos: Integrating Databases and Apache Kafka

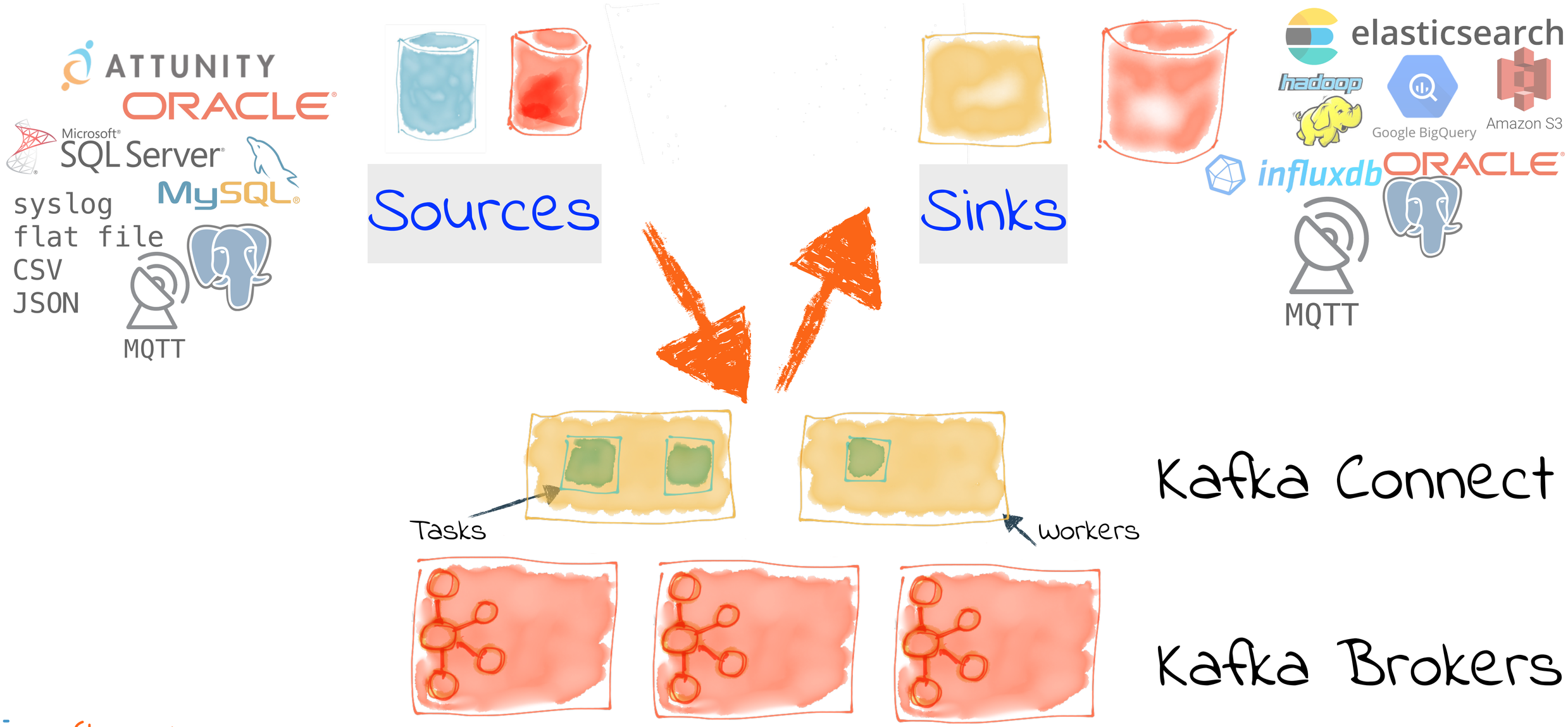
Streaming Integration with Kafka Connect



Streaming Integration with Kafka Connect

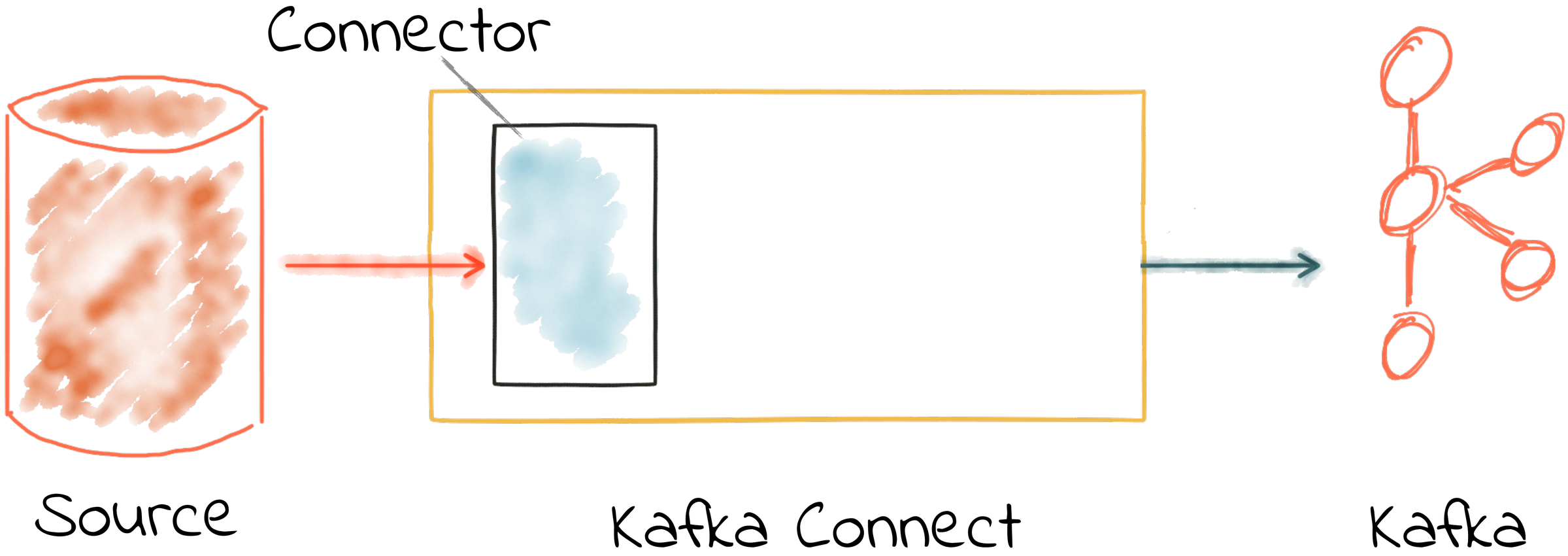


Streaming Integration with Kafka Connect

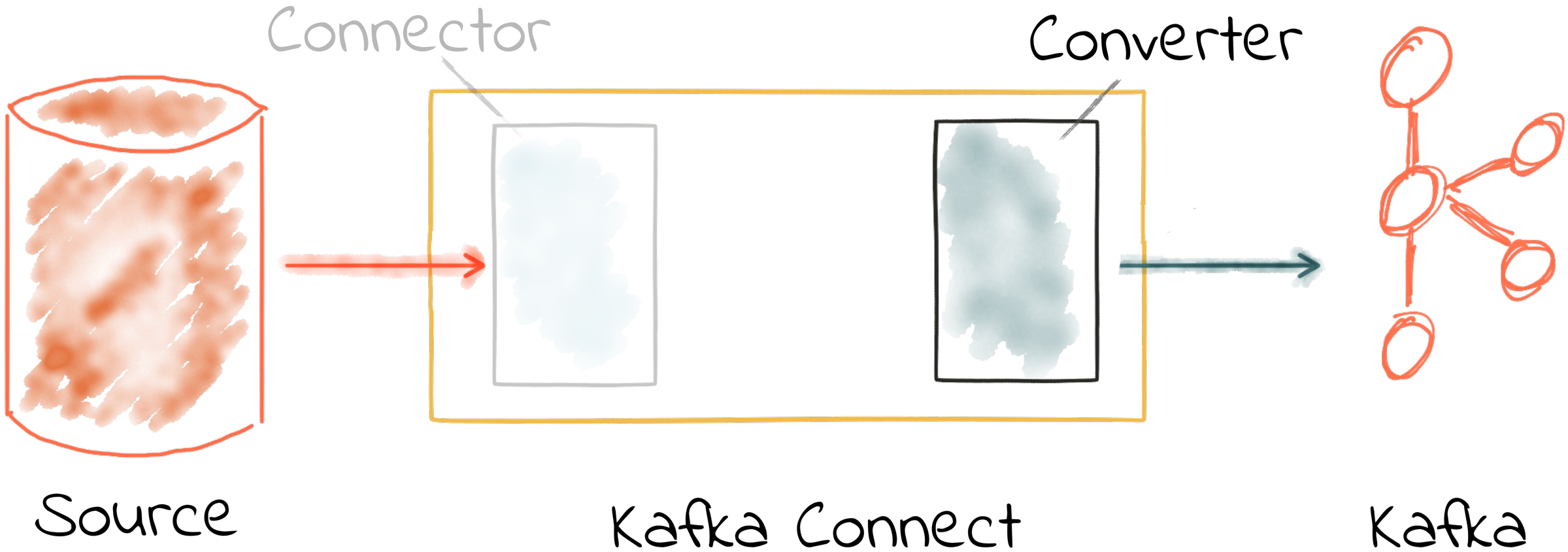


Kafka Connect basics

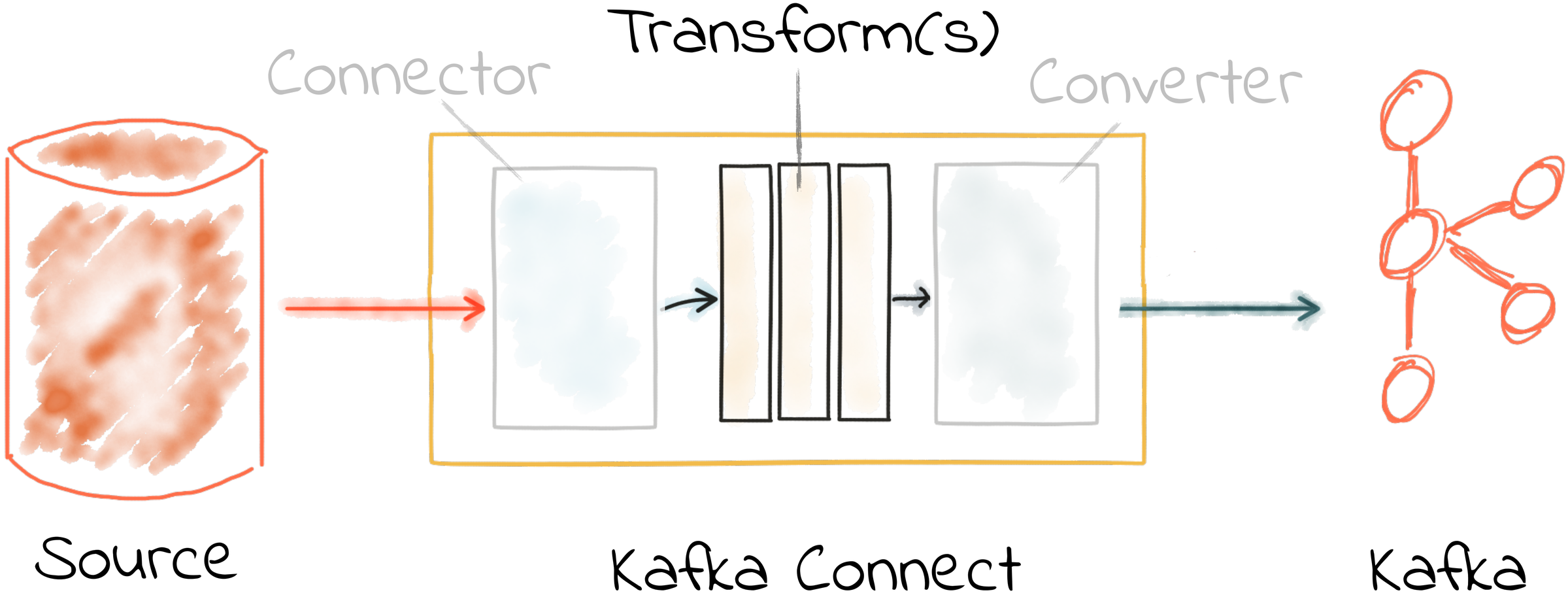
Connectors



Converters

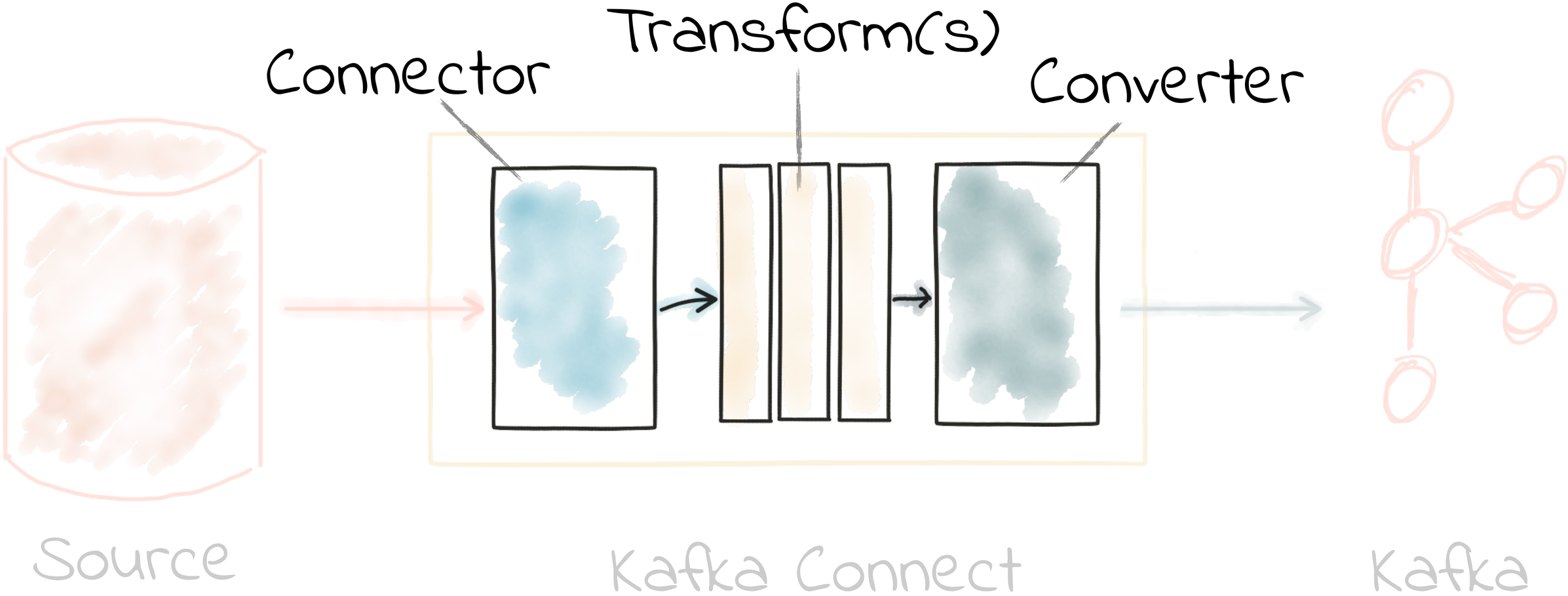


Single Message Transforms



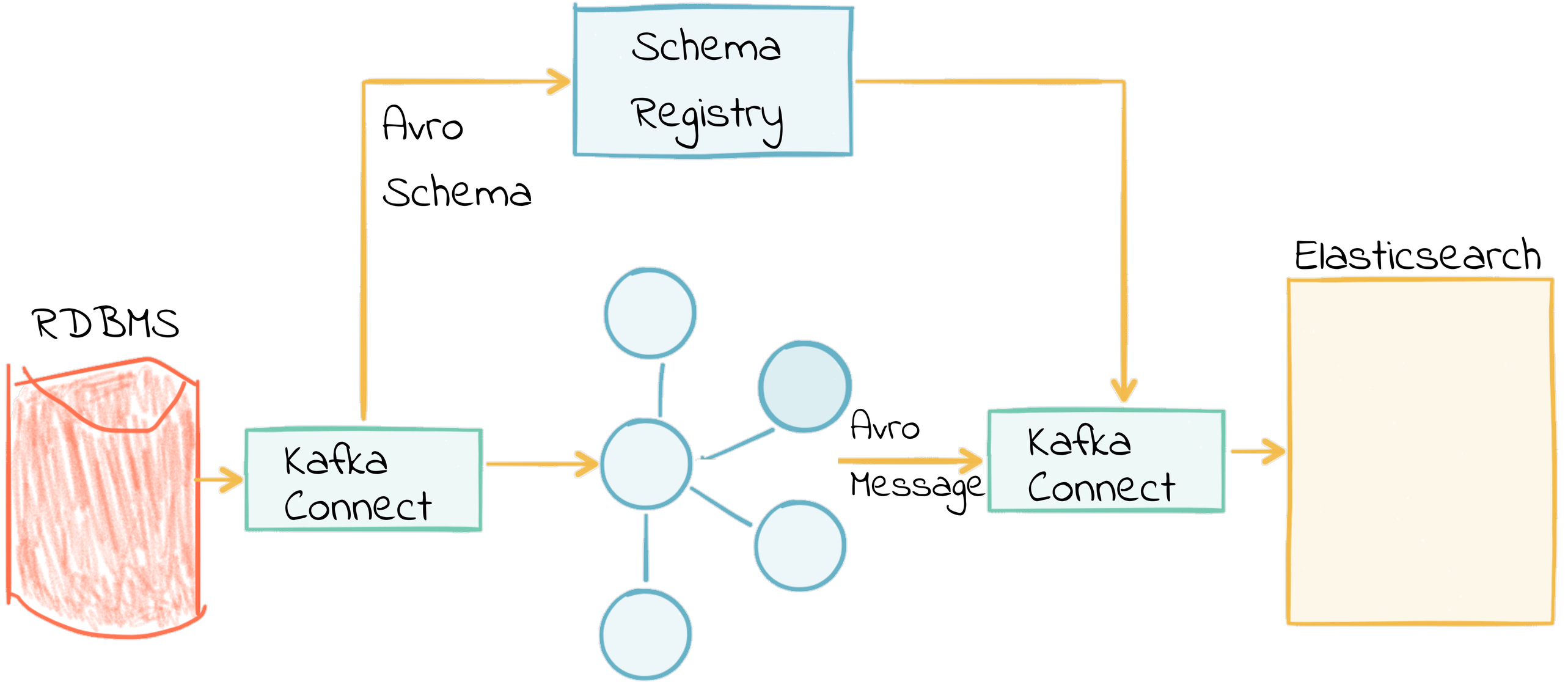
Extensible

Browse & download at hub.confluent.io

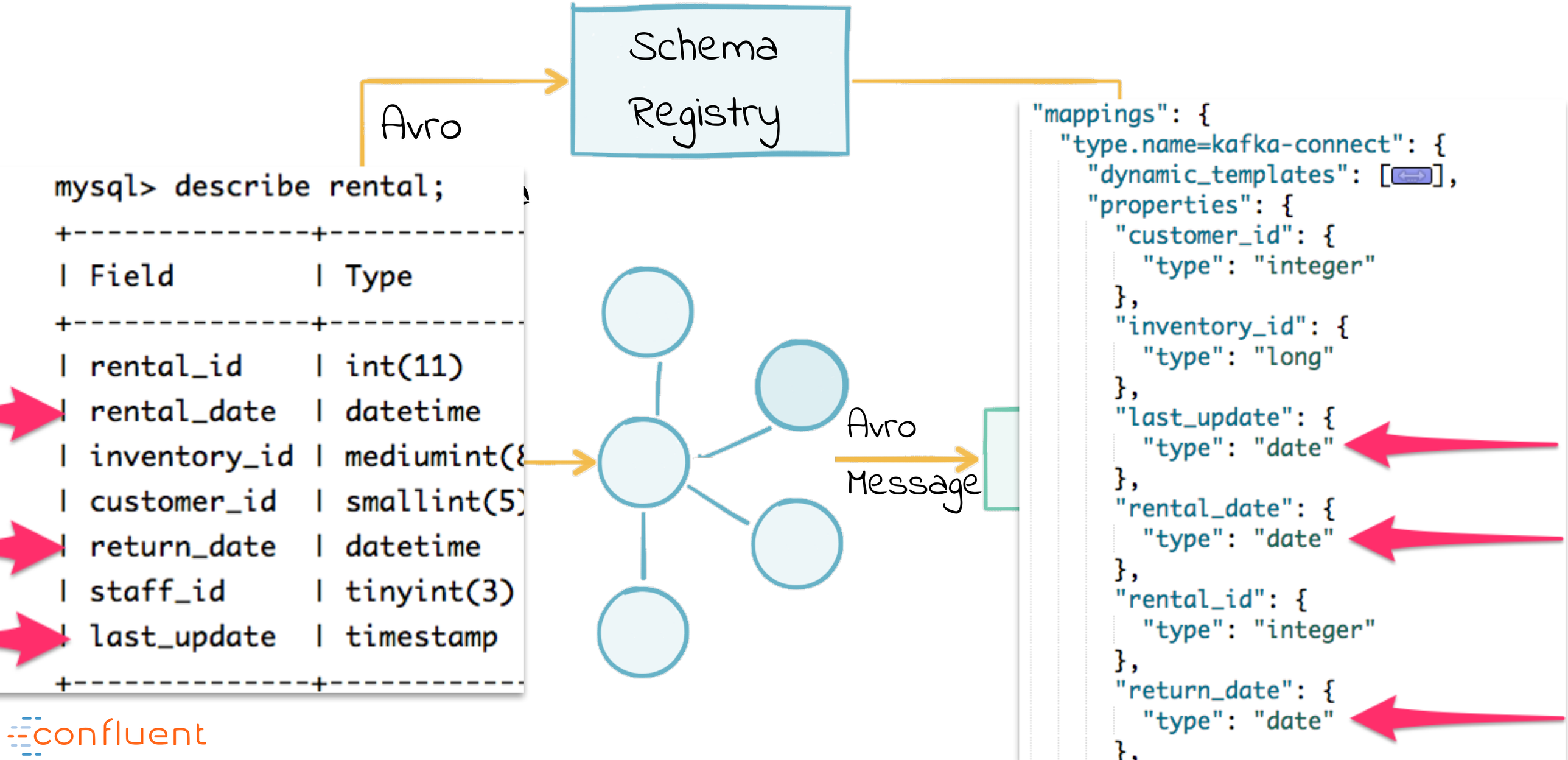


https://docs.confluent.io/current/connect/javadocs/

Kafka Connect + Schema Registry = WIN



Kafka Connect + Schema Registry = WIN



confluent

Change-Data-Capture (CDC)

- **CDC** is a generic term referring to *capturing changing data* typically from a RDBMS.
- Two general approaches:
  - **Query-based CDC**
  - **Log-based CDC**

There are other options including hacks with Triggers, Flashback etc but these are system and/or technology-specific.

Query-based CDC

- Use a database query to try and identify new & changed rows

```
SELECT * FROM my_table
WHERE col > <value of col last time we polled>
```

- Implemented with the open source **Kafka Connect JDBC connector**
  - Can import based on table names, schema, or bespoke SQL query
  - Incremental ingest driven through incrementing ID column and/or timestamp column

Log-based CDC

- Use the database's transaction log to identify every single change event
- Various CDC tools available that integrate with Apache Kafka (more of this later…)

```
Logdump 12 >pos 6636
Reading forward from RBA 6636
Logdump 13 >n
------------------------------------------------------------------
Hdr-Ind    :     E  (x45)     Partition  :     .  (x0c)
UndoFlag   :     .  (x00)     BeforeAfter:     A  (x41)
RecLength  :   256  (x0100)   IO Time    : 2016/09/06 11:59:23.000.589
IOType     :     5  (x05)     OrigNode   :   255  (xff)
TransInd   :     .  (x00)     FormatType :     R  (x52)
SyskeyLen  :     0  (x00)     Incomplete :     .  (x00)
AuditRBA   :         393      AuditPos   : 30266384
Continued  :     N  (x00)     RecCount   :     1  (x01)

2016/09/06 11:59:23.000.589 Insert               Len   256 RBA 6636
Name: ORCL.SOE.CUSTOMERS  (TDR Index: 3)
After  Image:                                             Partition 12   G  b
0000 000a 0000 0000 0000 0001 86a1 0001 000a 0000 | ..................
0006 616e 7477 616e 0002 000b 0000 0007 7361 6d70 | ..antwan........samp
736f 6e00 0300 0600 0000 0275 7300 0400 0b00 0000 | son........us.......
0741 4d45 5249 4341 0005 000a 0000 0000 0000 0000 | .AMERICA............
8980 0006 001d 0000 0019 616e 7477 616e 2e73 616d | ..........antwan.sam
7073 6f6e 406f 7261 636c 652e 636f 6d00 0700 0a00 | pson@oracle.com.....
0000 0000 0000 0000 9500 0800 1500 0032 3031 362d | ...............2016-
Column     0 (x0000), Len    10 (x000a)
0000 0000 0000 0001 86a1                          | ..........
Column     1 (x0001), Len    10 (x000a)
0000 0006 616e 7477 616e                          | ....antwan
Column     2 (x0002), Len    11 (x000b)
0000 0007 7361 6d70 736f 6e                       | ....sampson
Column     3 (x0003), Len     6 (x0006)
0000 0002 7573                                    | ....us
Column     4 (x0004), Len    11 (x000b)
0000 0007 414d 4552 4943 41                       | ....AMERICA
Column     5 (x0005), Len    10 (x000a)
0000 0000 0000 0000 8980                          | ..........
Column     6 (x0006), Len    29 (x001d)
0000 0019 616e 7477 616e 2e73 616d 7073 6f6e 406f | ....antwan.sampson@o
7261 636c 652e 636f 6d                            | racle.com
Column     7 (x0007), Len    10 (x000a)
0000 0000 0000 0000 0095                          | ..........
Column     8 (x0008), Len    21 (x0015)
```

Demo Time!

"which one should I use?"

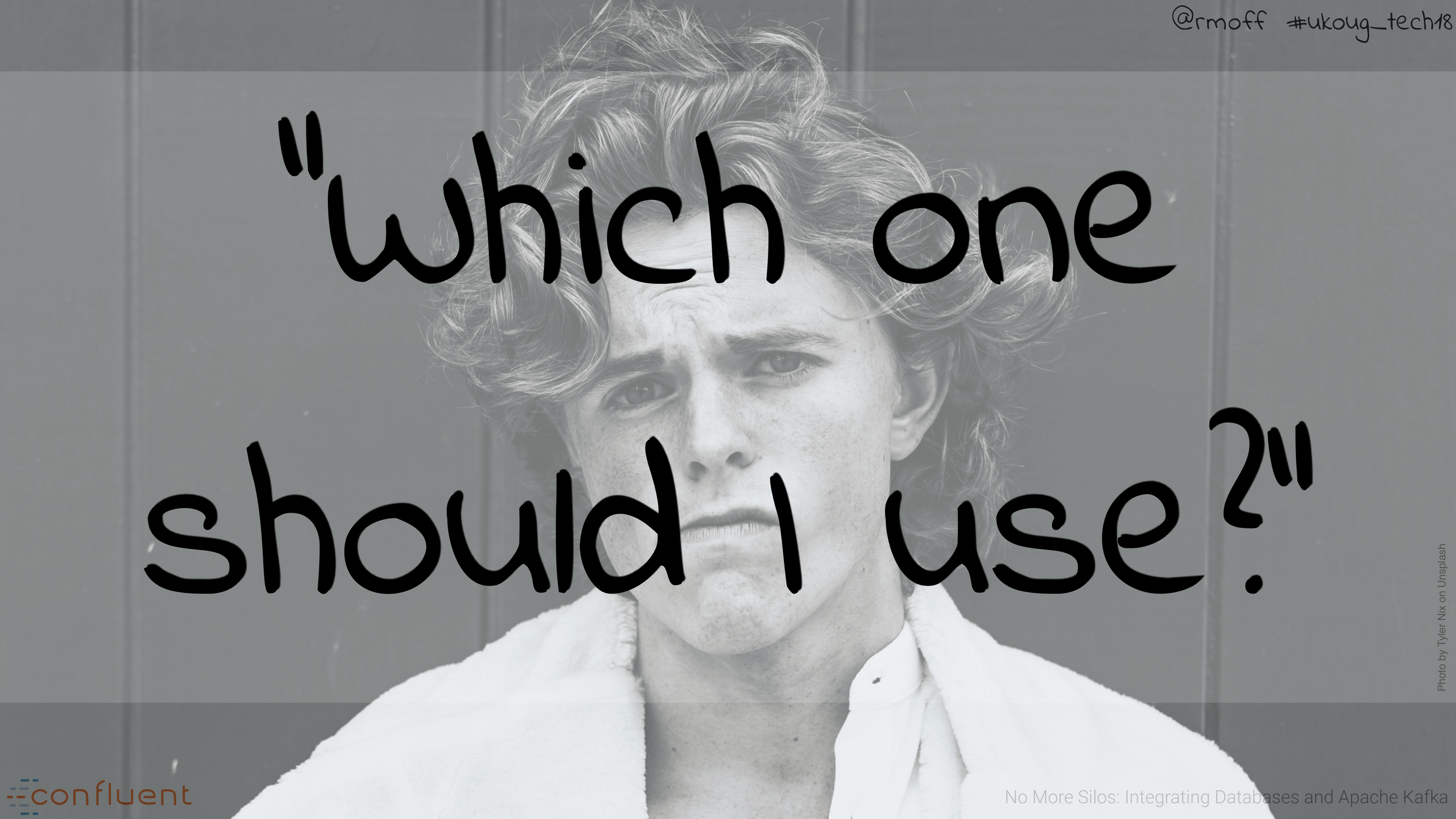


Photo by Tyler Nix on Unsplash

Photo by Trevor Cole on Unsplash

Query-based vs Log-based CDC

- **Query-based**

+Usually easier to setup, and requires fewer permissions

- Needs specific columns in source schema

- Impact of polling the DB (or higher latencies tradeoff)

- Can't track deletes, or multiple events between polling interval

Photo by Matese Fields on Unsplash

Query-based vs Log-based CDC

- **Log-based**
  - +Greater data fidelity
  - +Lower latency
  - +Lower impact on source
  - - More setup steps
  - - Higher system privileges required
  - - For propriatory databases, usually $$$

Read more: http://cnfl.io/kafka-cdc

Considerations for Integration into Apache Kafka

- Chucking data over the fence into a Kafka topic is not enough
- CDC tools should integrate with **standard ways of building data pipelines in Kafka**
  - Schema handling
  - Serialisation formats

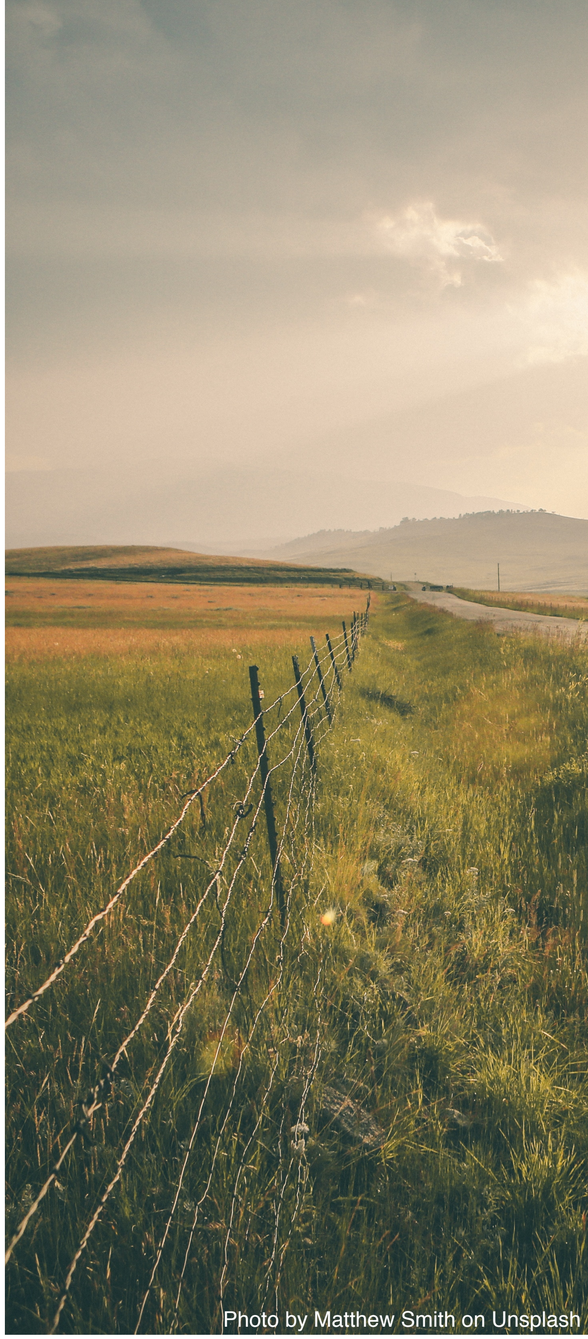


Photo by Matthew Smith on Unsplash

Oracle and Kafka integration

- **Oracle GoldenGate for Big Data**—Requires the OGGBD licence, not just OGG
- **Debezium**—Open source,Oracle support in Beta
  - currently uses XStream—which requires OGG licence
- **Attunity, IBM IIDR, HVR, SQData, StreamSets**—all offer commerical CDC integration into Kafka with support for Schema Registry
- **DBVisit Replicate**—no longer under development
- **JDBC Connector**—Open source, but not "true" CDC

Which Log-Based CDC Tool?

ⓘ *For query-based CDC, use the Confluent Kafka Connect JDBC connector*

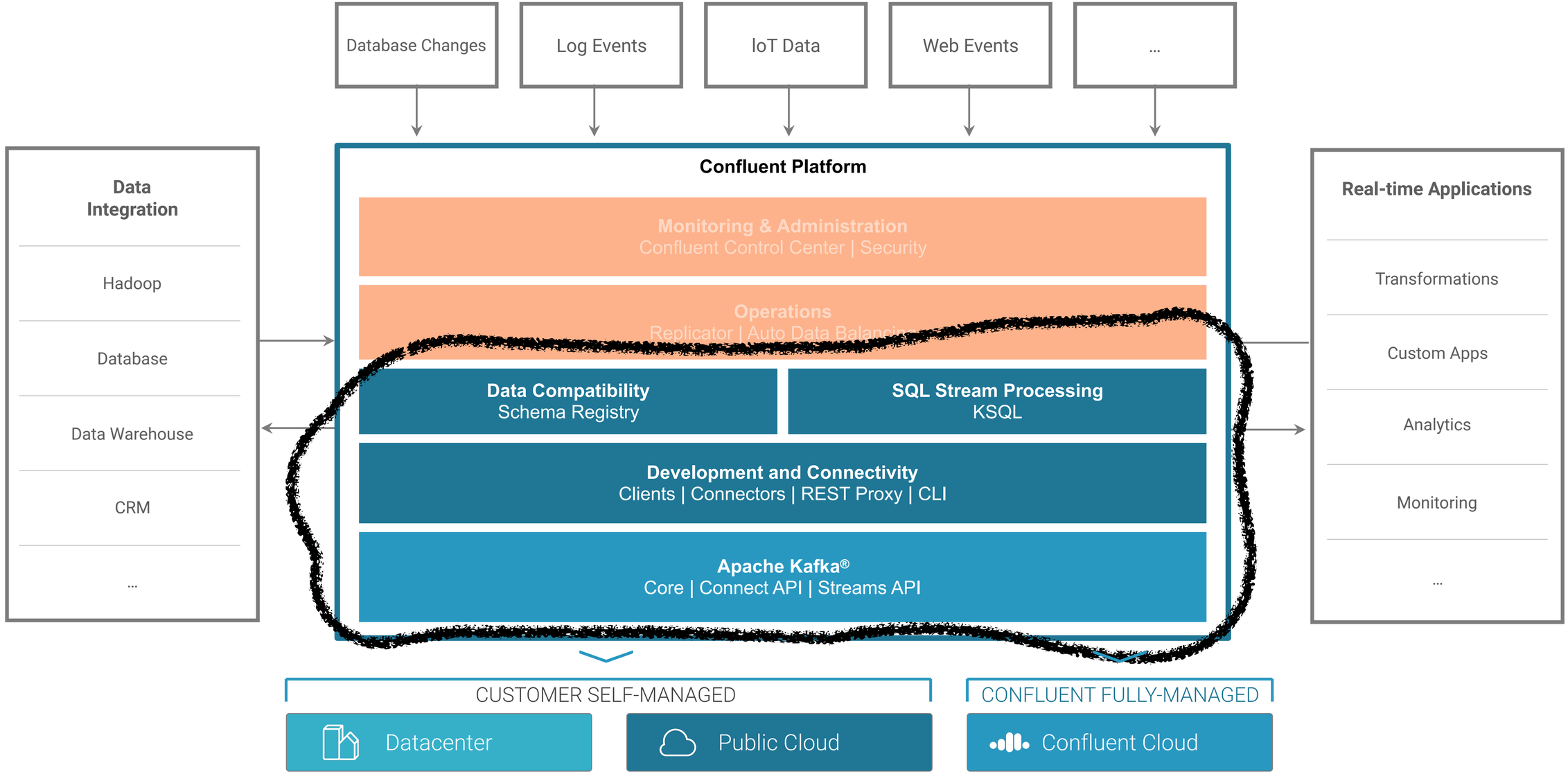
- Open Source RDBMS,
  e.g. **MySQL**, **PostgreSQL**
  - Debezium
  - (+ paid options)

- Mainframe
  e.g. **VSAM**, **IMS**
  - Attunity
  - SQData

- Proprietory RDBMS,
  e.g. **Oracle**, **MS SQL**
  - Oracle GoldenGate
    - Debezium + XStream
  - Attunity
  - IBM InfoSphere Data Replication (IIDR)
  - SQData
  - HVR

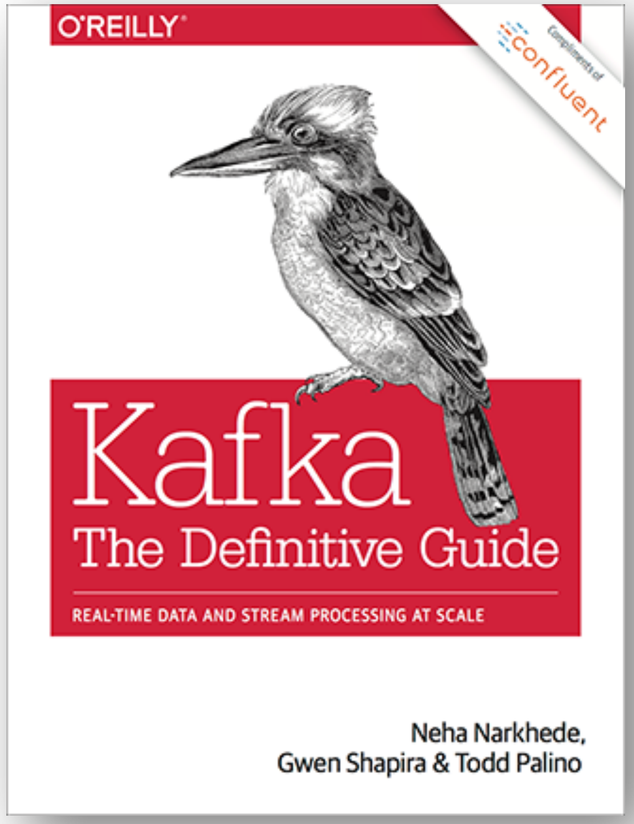
All these options integrate with Apache Kafka and Confluent Platform, including support for the Schema Registry

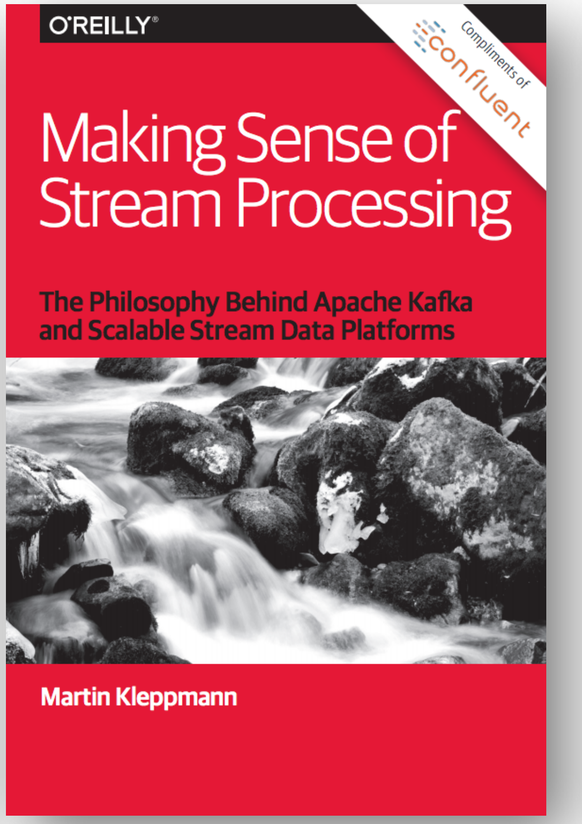
Confluent Open Source : Apache Kafka with a bunch of cool stuff! For free!

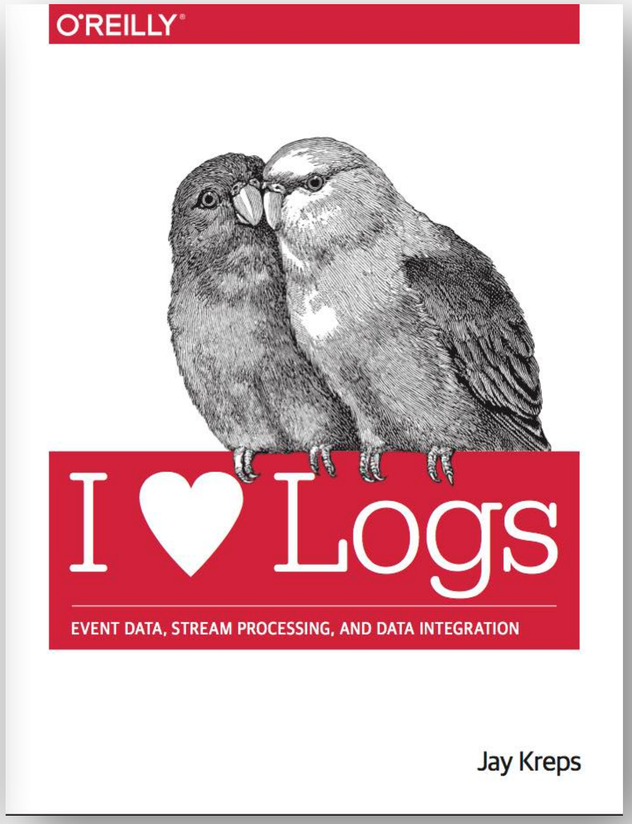


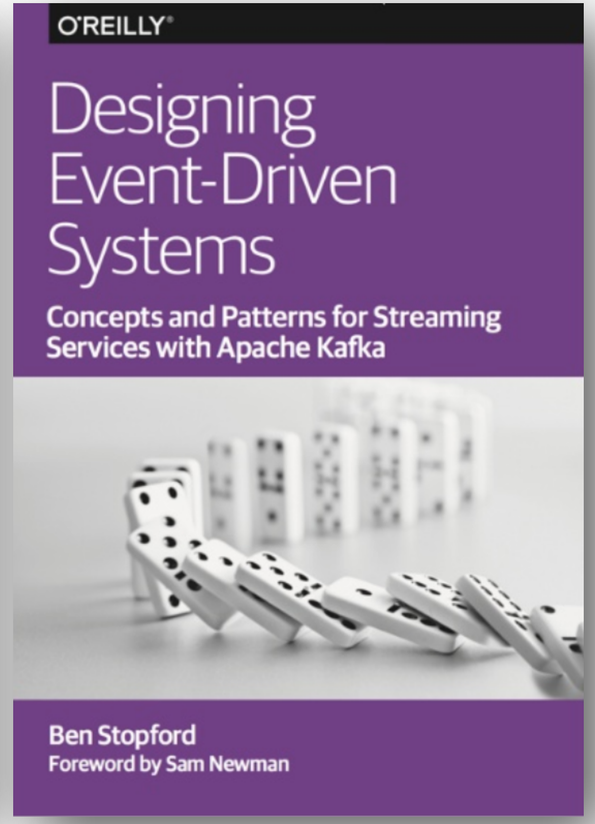
Free Books!

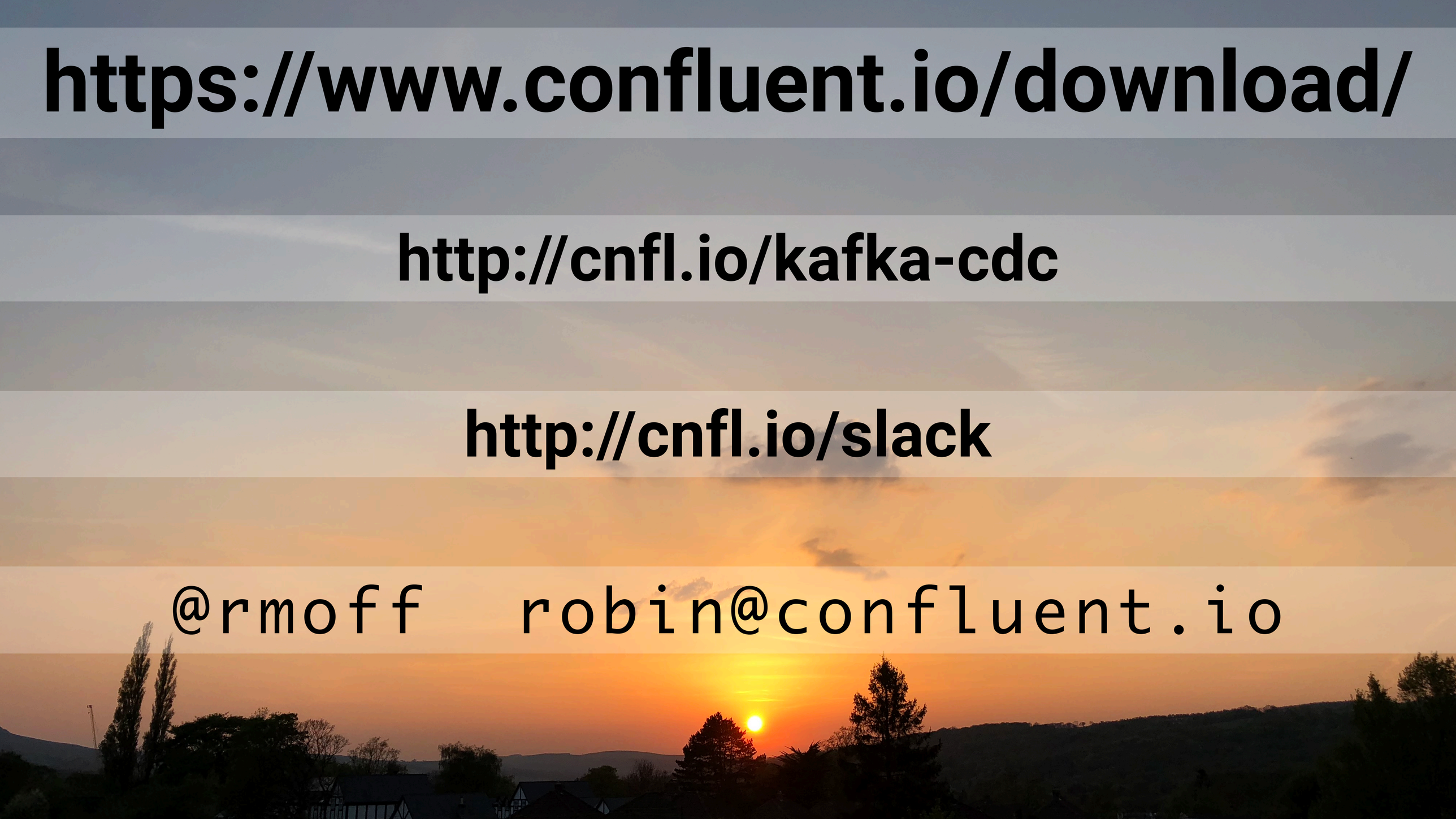
https://www.confluent.io/apache-kafka-stream-processing-book-bundle











https://www.confluent.io/download/

http://cnfl.io/kafka-cdc

http://cnfl.io/slack

@rmoff robin@confluent.io

#EOF